DOCUMENT RESUME

ED 408 853 FL 024 615

AUTHOR Daniel, Larry G.; Witta, E. Lea
TITLE Implications for Teaching Graduate Students Correct Terminology for Discussing Validity and Reliability Based on a Content Analysis of Three Social Science Measurement Journals.
PUB DATE Mar 97
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Graduate Students; Graduate Study; Higher Education; Language Role; Language Usage; *Research Methodology; *Research Problems; Scholarly Journals; *Social Sciences; *Test Reliability; *Test Validity; *Vocabulary Development

ABSTRACT

Although reliability and validity are characteristics of test data, social scientists often attribute reliability and validity erroneously to the tests themselves. To determine the extent to which this problem exists, 150 reliability and validity studies selected from 3 prominent social science measurement journals over a 3-year period were analyzed for common errors in terminology and categorized according to methodology types used in assessing reliability and validity. Results indicate over 50 percent of the articles contained more than one inappropriate statement concerning reliability or validity. It is suggested that professional journal reviewers and editors could improve research practice by catching and correcting a larger percentage of these errors. In the educational research classroom, it is recommended that teachers emphasize that reliability and validity are properties of data, model correct language about score characteristics while discussing reliability and validity in the presence of their students, and correct students' inappropriate use of language. Study data are appended. Contains 12 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 853

Running Head: DISCUSSING VALIDITY AND RELIABILITY

Implications for Teaching Graduate Students Correct Terminology for Discussing Validity and Reliability Based on a Content Analysis of Three Social Science Measurement Journals

Larry G. Daniel and E. Lea Witta

University of Southern Mississippi

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Larry Daniel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, March 24-28, 1997.

BEST COPY AVAILABLE

Abstract

Validity and reliability are characteristics of test *data*; however, researchers and professors of social science research often erroneously attribute validity and reliability to the tests themselves. To determine the extent to which this problem exists, as well as the degree to which various methodological concerns relative to the reporting of results of validity and reliability studies are present, in published research, 150 validity and reliability studies were selected from 3 prominent social science measurement journals over a 3-year period. These studies, taken from the 1992, 1993, and 1994 volumes of *Educational and Psychological Measurement*, *Psychological Assessment*, and *Journal of Psychoeducational Assessment* were reviewed for common errors in terminology and were categorized according to the types of methodologies employed in the assessment of validity and reliability. Implications of the findings for professors of educational research and measurement are offered.

Implications for Teaching Graduate Students Correct Terminology for Discussing Validity and Reliability Based on a Content Analysis of Three Social Science Measurement Journals

Measurement integrity is essential to the integrity of behavioral research. Consequently, the findings of any behavioral research study, no matter how well planned and executed, will be held suspect if information about the validity and reliability of the study's data is inadequate or missing. Simply put, any research hypothesis that includes variables operationally defined as test scores must be predicated upon sufficient evidence to substantiate the hypothesis that such test scores are valid and reliable (Messick, 1989; Pedhazur & Schmelkin, 1991), considering that a decision about the reliability and validity of test scores "is a special case of hypothesis testing" (ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1992, p. 1).

Even though training for research in the social science disciplines generally includes a considerable degree of attention to measurement integrity issues, the ways in which measurement integrity studies are conducted and reported do not always square with guidelines for best practice. In particular, critics within the scholarly community have, with a moderate degree of frequency, identified problems in the professional literature related to (a) use of inappropriate language in the reporting of results of analyses of the validity and reliability of test scores and (b) incidence of occurrence of methodological procedures that are in opposition to best practice in the reporting of validity and reliability estimates. These problems reflect negatively on the quality of the studies in which they occur and also have the potential to prompt misunderstandings about validity and reliability by those who read such studies.

Purpose

Considering the importance of developing accurate estimates of the validity and reliability of scores on tests selected or generated for use in social science research, the purpose of the present study was to gain an understanding of the degree to which various problems relative to misuse of language regarding measurement characteristics of test scores as well as various methodological concerns in the reporting of results of validity and reliability studies are present in recent social science measurement journals. Following a review of the literature related to the present study, we describe the methodology employed in the study, present results, and discuss the findings in light of their implications for teachers of educational research and measurement.

Review of Related Literature

In providing measurement integrity evidence to justify the use of scores from instruments utilized in a given study, many researchers are prone to report validity and reliability estimates derived using data collected with a given instrument in one or more previous studies. Although this practice is apropos to the purpose of establishing evidence for validity or reliability, it is not adequate in and of itself as a means for supporting the validity or reliability of the scores on the same instruments when used in a new study with a sample different from those sampled in previous studies. Data must be collected from a given sample in order to generate estimates of the validity or reliability of the data collected from that sample. Even when data of this type are collected for the sample utilized within a given study, the results of the analyses of those data can be adversely affected if the author (a) describes the data using inappropriate language or else (b)

reports results of validity and reliability studies that are not methodologically sound. Literature relative to these two problem areas will be reviewed herein.

Inappropriate Language Used Relative to Score Characteristics

Reliability and validity are always characteristics of test data, not the tests themselves. As Thompson (1994, p. 839) noted:

> . . .it becomes an oxymoron to speak of "the reliability [or validity] of the test" without considering to whom the test was administered or other facets of the measurement protocol. . . . [T]he same measure, when administered to more heterogeneous or to more homogeneous sets of subjects, will yield scores with differing reliability [and validity].

The *Standards for Educational and Psychological Testing* developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) in 1985 spoke directly to the common misperception that tests, in and of themselves, may be valid or reliable: "Validity. . .refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. . . . The inferences regarding specific uses of the test are validated, not the test itself" (p. 9--emphasis added). Linn and Gronlund (1995) echoed this sentiment, noting, "We sometimes speak of the 'validity of a test,' for the sake of convenience, but it is more correct to speak of the validity of the interpretation and use to be made from the results." Similarly, Wainer and Braun (1988) noted, "The 'validity of a test' is a misnomer" (p. 87), and Popham (1995) asserted, "Tests, themselves, do not possess validity" (p. 40).

Moreover, in an author guidelines editorial published in *Educational and Psychological Measurement*, Thompson (1994, p. 839) noted that loose use of language about test score characteristics is not only a sign of ignorance but also a potential antecedent to bad psychometric practice:

> One unfortunate feature of contemporary scholarly language is the usage of the statement "the test is reliable" or "the test is valid." Such language is both incorrect and deleterious in its effects on scholarly inquiry, particularly given the pernicious consequences that unconscious paradigmatic beliefs can exact. . . . Too few researchers act on a conscious recognition that *reliability [or validity] is a characteristic of scores or the data in hand.* (emphasis in original)

In the same editorial, Thompson noted a new editorial policy requiring authors submitting manuscripts to *Educational and Psychological Measurement* to use language (a) that is more technically correct (i.e., refers to reliability and validity of scores rather than tests) and (b) that would, therefore, reinforce better practice.

A related language use problem is the tendency of some researchers to overstate the case for the validity or reliability of the scores in their research studies. For example, some authors will claim that results of a given study "prove" or "demonstrate" that a given set of test scores (or worse yet, that the test itself) is valid or reliable. Statements of this type are erroneous for at least two reasons. First, as previously noted, measurement validity and reliability are specific to some particular use or interpretation of the data in hand (Linn & Gronlund, 1995). More

importantly, validity (and to somewhat of a lesser degree, reliability) is appropriately viewed as an evolving system of inferences rather than as a single set of data analytic procedures:

> . . .over time, the existing validity evidence becomes enhanced (or contravened) by new findings, and projections of potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions. Inevitably, then, validity is an evolving property and validation is a continuing process. (Messick, 1989, p. 13)

Hence, the results of any study may either confirm or disconfirm previous findings, but a study's results do not really "prove" or "demonstrate" validity.

Methodological Concerns in Reporting Validity and Reliability Estimates

There are at least two substantial methodological issues that may serve to convolute results of validity and reliability analyses. The first of these is the tendency of some researchers to report a statistical significance test along with a reliability or validity coefficient. Such tests typically evaluate the likelihood that the test scores are totally unreliable ($r = 0$). These statistical comparisons are virtually meaningless considering that for large coefficients, the null hypothesis may be rejected with an n as small as 5 (Thompson, 1994)! Moreover, since statistical significance is largely an artifact of sample size, a rather low coefficient may be statistically significant if the n is quite large (Huck & Cormier, 1996), possibly resulting in the careless conclusion that the coefficient signifies adequate reliability or validity. Besides these sample size arguments, an additional substantial argument against the use of significance testing for evaluating validity and reliability coefficients is that these coefficients by nature are sample

specific (e.g., validity and reliability are functions of the data in hand), and therefore, the coefficients would not be expected to be generalizable to a different sample drawn from the same population.

A second methodological issue that serves as cause for concern is the possibility of a set of data yielding a negative reliability coefficient. Obviously, these coefficients cannot be mathematically correct considering that a reliability coefficient indicates the proportion of the true score variance to the total observed variance in a set of scores. Even in the case in which a set of scores is completely unreliable, the coefficient should be no less than zero. Nevertheless, as illustrated by Krus and Helmstadter (1993), conventional formulae for estimating reliability coefficients can sometimes yield these counter-intuitive negative values. Although a prevailing logic is to simply set these values to zero, implying that the scales are completely unreliable, it is also possible that negative reliabilities may indicate that more than one construct is being measured and that those constructs are negatively correlated (Krus & Helmstadter, 1993). At any rate, negative reliabilities are quite problematic and therefore should not typically appear in published research studies.

Method

As previously noted, the present study sought to determine the degree to which the aforementioned language and methodology problems are manifest in articles appearing social science measurement journals. Three journals (i.e., *Educational and Psychological Measurement*, *Psychological Assessment*, and *Journal of Psychoeducational Assessment*) that regularly publish validity and reliability studies were selected as the source for the articles

reviewed. The volume years coinciding with the 1992, 1993, and 1994 calendar years were selected for each of the three journals. These volume years were included because they were relatively recent but also because the manuscripts selected for publication during this period of time were submitted prior to the author guidelines editorial included in *Educational and Psychological Measurement* (Thompson, 1994) which called for a moratorium on manuscripts which included the type of problems mentioned herein.

In selecting the articles to be sampled for review, the following procedures were utilized:

(1) All articles appearing in the "Validity Studies" section of *Educational and Psychological Measurement* over the 3-year period were considered. As shown in Table 1, a total of 190 articles from this journal were initially identified.

(2) All articles appearing in the main and "Brief Studies" sections of *Psychological Assessment* and the main section of *Journal of Psychoeducational Assessment* for the 3-year period were scanned to determine whether they were primarily reliability/validity studies or not. This process included reading all titles and abstracts and quickly perusing the articles' content. Based on this process, 85 *Psychological Assessment* and 25 *Journal of Psychoeducational Assessment* articles were initially identified (see Table 1).

(3) One hundred fifty articles were sampled from the 300 articles identified using procedures in steps (1) and (2) above. This sampling was done as follows. All 25 of the *Journal of Psychoeducational Assessment* articles were sampled since this subset was relatively small as compared to the subsets from the other two

journals. Fifty of the 85 *Psychological Assessment* articles were randomly sampled, and 75 of the 190 *Educational and Psychological Measurement* articles were sampled. This distribution made for a nice balance between the *EPM* and non-*EPM* articles, and created a relatively feasible coding load for the two raters.

Each of us read and coded 90 articles (30 were coded by both raters) using the rating form shown in the appendix, leaving 120 articles that were uniquely coded by a single rater (60 articles per rater). Fifteen of the "double coded" articles were selected from each rater yielding a grand total of 150 articles for this analysis. The rating form allowed for articles to be coded on each of the following criteria (see Appendix A):

(1) Was erroneous language implying the validity or reliability of a test used

(a) in the title,

(b) in the abstract,

(c) in the study?

(2) Were statistical significance tests reported along with validity or reliability coefficients?

(3) Was erroneous language used suggesting that findings had proven or demonstrated the validity/reliability of data/tests?

(4) What type(s) of reliability evidence was(were) provided (content, predictive, concurrent, construct, convergent, discriminant)?

(5) If construct validity evidence was provided, what statistical procedure(s) did the author employ (exploratory factor analysis, confirmatory factor analysis, multitrait matrices, multitrait-multimethod matrices)?

(5) What method(s) was(were) used to assess validity?

Results and Discussion

Inter-rater reliability was determined by both researchers coding 30 articles. If an inappropriate use of validity/reliability (see Appendix A) was noted, this variable was coded as 1. Otherwise the variable was coded as 0. These responses were summed to form a rating score and entered into a two-way ANOVA in SPSS/PC+. We calculated the inter-rater generalizability coefficient using the formula,

$$\rho_1^2 = \frac{MS_p - MS_r}{MS_p}$$

where MS_p represents the mean square article and MS_r is mean square error (Crocker & Algina, 1986, p 167). Inter-rater reliability was .83 for the raters' codes across the selected studies. Actual agreement/disagreement of the raters for each inappropriate category is displayed in Table 2. Raters agreed on 240 of the possible 270 responses, yielding a rater agreement percentage of 89%.

Almost 51% (76) of the 150 studies analyzed referred to the test (or scale) as valid (see Table 3). In addition, 43% (65) of the studies refer to the reliability of the test. Inappropriate language examples basically took one of two major syntactical forms: (a) "the _____ test is reliable (and/or valid)"; or (b) "reliability (and/or validity) of the _____ test." No study,

however, reported a negative reliability. Only one study referred to the statistical analysis as "proving" or "demonstrating" reliability and validity.

When evaluated by journal (see Table 4) by year, over 50% of the articles contained inappropriate language falling into more than one category. This finding implies that these journals regularly accepted articles using this terminology during the three-year time frame under study. It was generally obvious in reviewing the studies that researchers know a test is neither valid nor reliable. Most studies using "the test was valid ..." or "the scale was reliable . . ." then followed this statement with a recommendation that future research determine the validity/reliability in other settings or with other populations. The difficulty is not that experienced researchers do not know that it is scores that are valid or reliable, but that fledgling researchers may interpret a reported validity or reliability coefficient "of the" test as the test's validity or reliability and see no need to examine their sample. Indeed, some students appear to believe that once evidence for validity or reliability is established, it applies to that test in all situations with all populations. They (the students) think validity has become a characteristic of the test that does NOT vary.

Eighty-one percent of the 79 studies testing reliability used internal consistency estimates (see Table 5). This was followed by 33% in which test-retest estimates were utilized. A negligible number of studies reported use of inter-rater or alternate forms coefficients, and intra-rater reliability. Obviously Cronbach's alpha is much more frequently used than other methods of estimating reliability. Many studies utilized more than one type of reliability estimate.

Of the 92 studies estimating validity, 55% used construct validity procedures. Some of these used exploratory factor analysis, some used confirmatory factor analysis, and some used both procedures and/or others. Multi-trait, multi-trait multi-method and content validity estimates were the least frequently reported.

Conclusions/Recommendations

Over 50% of the journal articles reviewed contained more than one inappropriate statement concerning reliability or validity. In fact, 76 articles (50.7%) suggested the test/scale was valid while 65 (43.3%) suggested the test/scale was reliable. Even if the raters had a 10% error rate, would a 30% rate of use of inappropriate language be acceptable? Professional journal reviewers and editors could serve to improve research practice by catching and correcting a larger percentage of these errors. Certainly, the revised editorial policies adopted by *Educational and Psychological Measurement* (Thompson, 1994) have the potential to make a positive impact on language usage. This policy, at minimum, will make the reviewers and editors of that publication more aware of the problem. We contend that it would be wise for other social science measurement journals to adopt and implement similar editorial policies. Hopefully, a trend in this direction would eventually encourage editorial boards of non-measurement journals to also adopt these policies.

Within educational research classes our alternatives are limited. Clearly, the findings of the present study demonstrate the need for professors to emphasize that reliability and validity are properties of data, NOT tests. In addition to professors modeling correct language about score characteristics while discussing validity and reliability in the presence of their students,

correction of students' inappropriate usages of language might also enhance the students' likelihood of internalizing a sense of correctness of language. However, even though students may try to satisfy the professor while in class by using correct terminology, if exposure to prominent measurement journals indicates validity and reliability apply to the test, who is the student to believe?

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.

ERIC Clearinghouse on Tests, Measurement, and Evaluation. (1992). A guide to test validation. Washington, DC: American Institutes for Research.

Huck, S. W., & Cormier, W. H. (1996). Reading statistics and research (2nd ed.). New York: HarperCollins.

Krus, D. J., & Helmstadter, G. C. (1993). The problem of negative reliabilities. Educational and Psychological Measurement, 53, 643-650.

Linn, R. L., & Gronlund, N. E. (1995). Measurement and assessment in teaching (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Popham, W. J. (1995). Classroom assessment: What teachers need to know. Boston: Allyn and Bacon.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Wainer, H., & Braun, H. I. (1988). Test validity. Hillsdale, NJ: Erlbaum.

Table 1

Breakdown of Initially Identified and Sampled Articles by Journal and Publication Year

	Psychoeducational Assessment		Psychological Assessment		Educational & Psychological Measurement	
	Count	%	Count	%	Count	%
Initially Identified[a]						
1992	10	40.0	33	38.8	65	34.2
1993	10	40.0	32	37.6	64	33.7
1994	5	20.0	20	23.5	61	32.1
Total	25	100.0	85	100.0	190	100.0
Sampled[b]						
1992	10	40.0	17	34.0	23	30.7
1993	10	40.0	16	32.0	27	36.0
1994	5	20.0	17	34.0	25	33.3
Total	25	100.0	50	100.0	75	100.0

Note. [a]Percentage breakdown across titles: *EPM*--63%, *PA*--28%, *JPA*--8%. Total n = 300.
[b]Percentage breakdown across titles: *EPM*--50%, *PA*--33%, *JPA*--17%. Total n = 150.

Table 2

Percentage of Agreement/Disagreement for Inter-rater Reliability by Inappropriate Category

	Agree		Disagree	
	Count	%	Count	%
Title				
Test reliable/valid	26	87%	4	13%
Abstract				
Test is Reliable	27	90%	3	10%
Test is Valid	26	87%	4	13%
Study				
Negative Reliability	30	100%	0	0%
Prove	30	100%	0	0%
Test is Reliable	25	83%	5	17%
p value Reliability	29	97%	1	3%
Test is Valid	23	77%	7	23%
p value Validity	24	80%	6	20%
TOTALS	240	89%	30	11%

Note. 30 articles coded by both raters

Table 3

Frequency of Inappropriate Terminology by Category

Category	Count	%
Title		
Test reliable/valid	42	28.0%
Abstract		
Test is Reliable	40	26.7%
Test is Valid	51	34.0%
Study		
Negative Reliability	0	0%
Prove	1	.7%
Test is Reliable	65	43.3%
p value Reliability	6	4.0%
Test is Valid	76	50.7%
p value Validity	38	25.3%

Note. Sample size = 150.

Table 4
Frequency of Inappropriate Terminology by Journal and Year

	Year Reviewed					
Inappropriate	1992		1993		1994	
Validity/Reliability	Count	%	Count	%	Count	%
Psychoeducational Assessment						
None	2	20%	2	20%	1	20%
One Instance	3	30%	3	30%	0	0%
>1 Instance	5	50%	5	50%	4	80%
Psychological Assessment						
None	5	29%	3	19%	5	29%
One Instance	1	6%	5	31%	2	12%
>1 Instance	11	65%	8	50%	10	59%
Educational & Psychological Measurement						
None	4	17%	4	15%	12	48%
One Instance	7	30%	8	30%	3	12%
>1 Instance	12	52%	15	56%	10	40%
For All Journals						
None	11	22%	9	17%	18	38%
One Instance	11	22%	16	30%	5	11%
>1 Instance	28	56%	28	53%	24	51%

Note. All percentages are calculated per year per journal.

Table 5

Frequency of Methods used to Determine Reliability or Validity

Method	Count	%
Validity (n = 92)		
Content	3	3.3%
Predictive	24	26.1%
Concurrent	28	30.4%
Construct	51	55.4%
Exploratory Factor	34	37.0%
Confirmatory Factor	19	20.7%
Multitrait	1	1.1%
Multitrait-Multimethod	3	3.3%
Convergent/Discriminant	26	28.3%
Reliability (n = 79)		
Test-retest	26	32.9%
Equivalent Forms	1	1.3%
Split-Half	2	2.5%
Internal Consistency	64	81.0%
Inter-rater	5	6.3%
Intra-rater	0	100%

Appendix A

Article No _____ Reviewer Larry ___ Lea ___

****

Inappropriate Terminology - Does the article report:

Title

Test is reliable/valid _____

Abstract

Test is valid _____

Test is reliable _____

Study

Test is reliable _____

p value _____

Test is valid _____

p value _____

Use of PROVE _____

Negative Reliability _____

****

Reliability Form:

Test-Retest _____ Equivalent Forms _____

Split-Half _____ Internal Consistency _____

Inter-rater _____ Intra-rater _____

****

Validity Form:

Content ____ Predictive ____

Concurrent ____ Construct ____

Exploratory Factor ____ Confirmatory Factor ____

Multitrait-Multimethod ____ Convergent/Discriminant ____

****

Other: